Can web-based statistic services be trusted?

Blekinge Institute of Technology, Ronneby, Sweden
Bachelor thesis in Computer Science
C-level, 10 p

Authors:	Supervisor:	Examiner:
Sara Birkestedt Andreas Hansson	Jenny Lundberg	Guohua Bai

Abstract

Title Can web-based statistic services be trusted?

Authors Sara Birkestedt and Andreas Hansson

Supervisor Jenny Lundberg

Examiner Guohua Bai

Introduction A large number of statistic services exist today, which shows that there is a great interest in knowing more about the visitors on a web site. But how reliable is the result the services are giving? At least three different systems for traffic measurement exist in Sweden today because of a disagreement about which measurement systems that show the correct figures.

Hypothesis Web-based statistic services do not show an accurate result.

Purpose The purpose of the thesis is to find out how accurate the web-based statistic services are regarding unique visitors and number of pages viewed. Our hope is that this thesis will bring more knowledge about the different statistic services that exists today and the problems surrounding them. We will also draw attention to the importance of knowing how your statistic software works to be able to interpret the results correctly.

Method To investigate this, we chose to do practical tests on a selection of web-based statistic services. The services registered the traffic from the same web site during a test period. During the same period a control program registered the same things and stored the result in a database. In addition to the test, we chose to do an interview with a person working with web statistics, to get a better understanding of the area.

Conclusion Our investigation showed that there are big differences between the results from the web-based statistic services in the test and that none of them showed an accurate result, neither for the total number of page views nor unique visitors. This led us to the conclusion that web-based statistic services do not show an accurate result, which verifies our hypothesis. Also the interview confirmed that there is a problem with measuring web statistics.

Key words Web statistics, Internet measurements, Web metrics

Table of contents

1 Introduction

1.1 The problem area

Internet today has over 580 millions of visitors each day, and the number is increasing (Hallström, 2003). But how many of these people find their way to your site and what are they doing there? The number of services available for getting web statistics today show that there is an interest in this. But how reliable is the result the services are giving? That is what we have been trying to find out in this thesis.

Our interest in this subject was triggered when we read a bachelor thesis from Blekinge Institute of Technology (Elofsson & Larsson, 2002) that said that the result from statistical services often differed. Since we had some own experience from statistic services, this made us want to know more about this area. When we started to look into this, it showed to be several problems involved in the area of web statistics, and one example is that there is no standard today for how to measure web traffic. After a quick search on the Internet we found several pages that brought up this problem.

We have also been in contact with a person with knowledge in this subject, the system administrator at the web company NoName4us AB, Robin Ericsson, and asked him if he was aware of that the statistics might not be accurate. His answer was "*Yes, I am aware of the problem with getting correct statistics. On larger sites we use more than one system to verify the statistics.*" (Ericsson, 2003). This confirmed our theory that there is a problem with web-based statistic services.

Torun Bjurman, who is describing the problem in an article in Computer Sweden, says that there is a great disagreement about which of the web traffic measurement systems that show the correct figures. As an example she mentions that the web page www.aftonbladet.se tested two different software systems. The first software, Jupiter MMXI, measured 1.8 million unique visitors. The second software, Website Index, measured 3.6 million unique visitors for the same time period. (Bjurman, 2002) This shows that the software programs measures very differently, so how can you tell which of the figures that is correct?

At least three different systems for traffic measurement exist in Sweden today because of the problems with deciding which system that should be used. This has made the figures from the different systems impossible to compare with each other. The Swedish organisation KIA is currently working on recommendations for a new standard. (Myrén, 2003)

Just by looking at the first sources found when we started our research, our suspicions were confirmed. There seem to be issues to handle in the web measurement area, as what to measure and how to measure it.

To explain abbreviations and certain expressions used in this thesis, we have put together a small glossary. This is placed as an appendix at the end of the thesis.

1.2 Hypothesis

The hypothesis we have formulated during our work with this thesis is:

Web-based statistic services do not show an accurate result.

1.3 Research questions

To help us in our work with the thesis, we formulated these research questions:

- Can web-based statistic services calculate unique visitors correctly?
- Is it possible to use web-based statistic services to estimate visiting trends?
- Is there one kind of system that is more reliable than others when measuring unique visitors and page views?

1.4 Goal

The goal of this thesis was to find out if web-based statistic services are not reliable, i.e. the statistic services do not show the correct data about the visitors of the web site.

1.5 Purpose

With this thesis we wanted to find out how accurate the web-based statistic services are regarding unique visitors and number of pages viewed. We also wanted to inform users and future users of web-based statistic services about the problems in this area, to make them aware of that these services may not be measuring correctly and therefore not reliable.

Our hope is that this thesis will bring more knowledge about the different statistic services that exists today. We will also draw attention to the importance of knowing how your statistic software works to be able to interpret the results correctly.

1.6 Delimitations

In the theoretical part of our thesis, we have focused on what to measure. This is a very broad subject that includes not only what to measure but when to measure it, how to measure etc. We decided to look deeper into the web metrics, because this should be the first step – you have to know what to measure before you can decide how to do it.

When it comes to the practical part, due to the restrictions in time and resources we have delimited our work into only comparing the result from the different services, by measuring unique visitors and page views. This means that we have not looked into usability issues, price, features etc.

There are a lot of different web-based statistic services available today and we did not have the possibility to test all of the different software systems. Therefore we decided to choose five web-based statistic services for the test.

1.7 Target group

This thesis is directed to commercial and non-commercial web site owners and web developers that are interested in knowing more about the visitors on their sites, what information you can get about them and how to get it.

It may also be of interest for people interested in computer science, especially in the area of web/Internet.

1.8 Evaluation of used material

We have had difficulties finding published books in this area. Despite the web measurements problems that have existed for several years, there do not seem to be very much written.

There is a large commercial interest in this area, and it seems to be considerably larger than the academic. A large quantity of web sites takes up this matter, for example all companies that are providing services for web measurement. However, we have had difficulties finding any academic sources. Many of the experts in this field seem to be in commercial companies.

Most of our sources have been found through Blekinge Institute of Technology's services for information search, including the service ELIN which makes many electronic sources available. We considered articles from these services to be reliable, since the Institute library has selected the magazines and papers, but of course we have checked all sources we have used ourselves. When we have found electronically published sources of interest, we have tried to get hold of the actual printed magazine and used that instead. We have also found a lot of articles from different magazines, and we have tried only to choose from reliable and serious magazines and papers.

All material from WWW has been evaluated with thoroughness. The information about the services used in the test is taken from the web sites of the companies. This information should be treated with some scepticism, since there are commercial interests involved.

The freshness of the articles is important, since the Internet is an area of fast growth. However, this does not seem to apply to traffic analysis to the same extent, since several articles from different points of time seem to bring up the same problems. But with this in mind, we have tried to use newer sources as long as it has been possible to get the information of current interest.

2 Background

2.1 The need for web statistic analyses

For a long time the number of hits was the measure of how popular or "cool" a web site was. Soon people discovered that it was more interesting to know what the visitors where doing while they were visiting your site; what pages did they stay a long time on, when did they decide to leave etc. This lead to the development of site analysis programs which today can be very sophisticated and provides a lot of information (Morris, 1998). This development also applies to the field of e-commerce, where in the early days web site traffic analysis meant nothing more than to install a counter on your web site and run a simple statistics program on your log file. But business leaders soon discovered that this simple hit count was both inaccurate and not detailed enough for marketing purposes. (Coopee, 2000)

A lot of people today have their own non-commercial web site and they are often using free web hotels. Spray, Passagen, Home and Angelfire are just a few that are available. The Swedish service Passagen has over 100 000 web pages today according to information on their web site (http://www.passagen.se/funktioner/hjalp/medlemskap/index.shtml). People using these free web hotels do not have access to the web logs and therefore cannot see information about the visitors on their pages. To be able to see web statistics for their pages they have the possibility to use a service that will keep track of their visitors for them. There are a number of services like this today, both free and for a fee.

With web statistics you get a good measurement on how your site is increasing its popularity, perhaps according to changes you have made on the site. You can also see how heavy the traffic is on your site to be able to avoid a crash. It also gives you a possibility to optimise your site to your visitors' web browsers etc. Information about your visitors can be useful not only for the Webmaster but also the technicians, programmers and maybe most important, to give input to the marketing section and the web editor, depending on the purpose of the web site. (http://www.publiceringsverktyg.info/1742.html)

With proper user tracking you can personalize your site to your customers and thereby boost user loyalty and increase selling opportunities. Knowing more about your visitors can have huge impact on design and usability of your site. The services can be used to increase brand awareness, user loyalty and sales (Coopee, 2000). This shows that web statistics can be a powerful tool that can give companies great advantages in several areas, if it is used in a correct way.

Companies are under increasing pressure to document their web site's value, and many are investing in the different solutions that exist today. By 2006 the annual spending on site analytics will reach $1 billion according to Jupiter Research, and ASP-based solutions will account for 29% of that (Patton, 2002). Patton here says that the use of web

statistics will increase in the future, and together with the advantages the services offer; it is an important choice for a company.

2.2 An analysis of the measurement problems

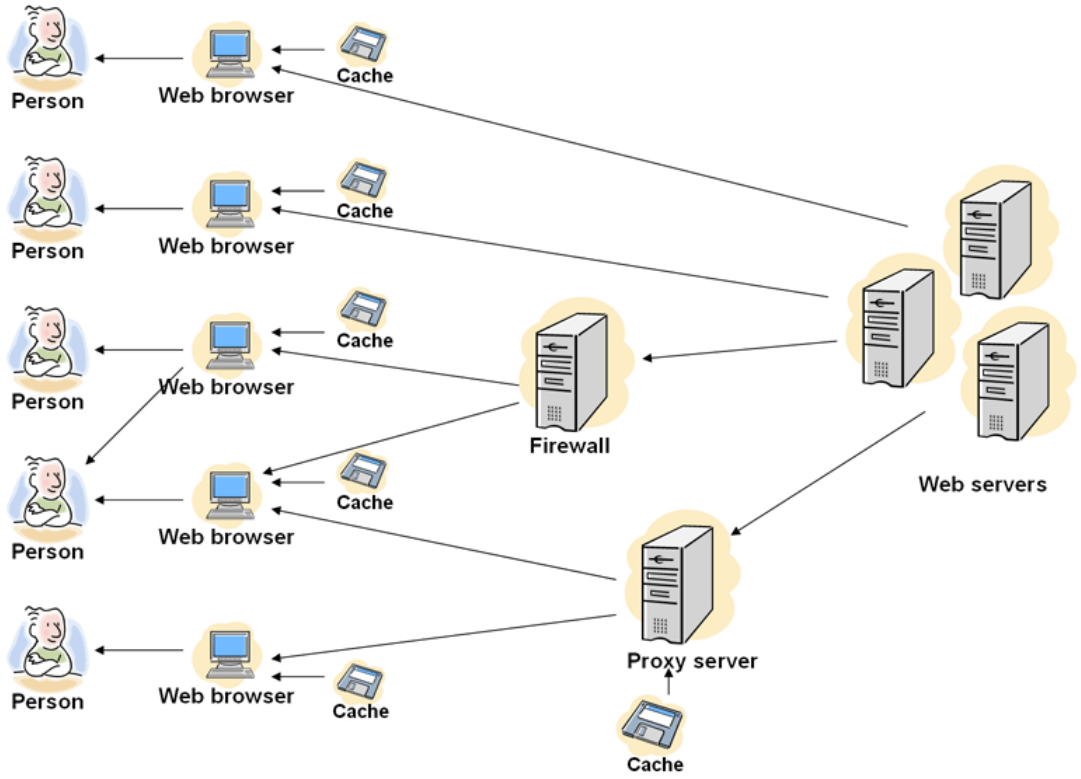
To understand the problems surrounding the collection and analysis of web statistics, it is important to understand some things about the technique behind the web, what components does the net consist of and how do they impact the statistics?

In this section we will try to analyse and describe the problems that exist and why they exist. As an example we will use a company, where the manager wants to have some information about the company's web site. The questions he has are:

- How many visitors did the web site have last year?
- Where did the visitors come from?
- How many pages did every visitor look at on an average?
- Which pages were most popular?

So, why are these questions difficult to answer? To understand this we will start to describe the web; what is the web and does it impact on the statistics? The web is built by the protocol HTTP, Hyper Text Transfer Protocol. (Saarelainen, 1996, p 278) HTTP is stateless, which means that the web server does not save any connections. The web server and web browser make a new connection for every file that is sent. (Fletcher, 2002) This lack of sessions means that visits does not really exist; they have to be created by using cookies. Temporary cookies can create sessions (visits). Permanent cookies can identify a web browser, even behind proxy servers and firewalls.

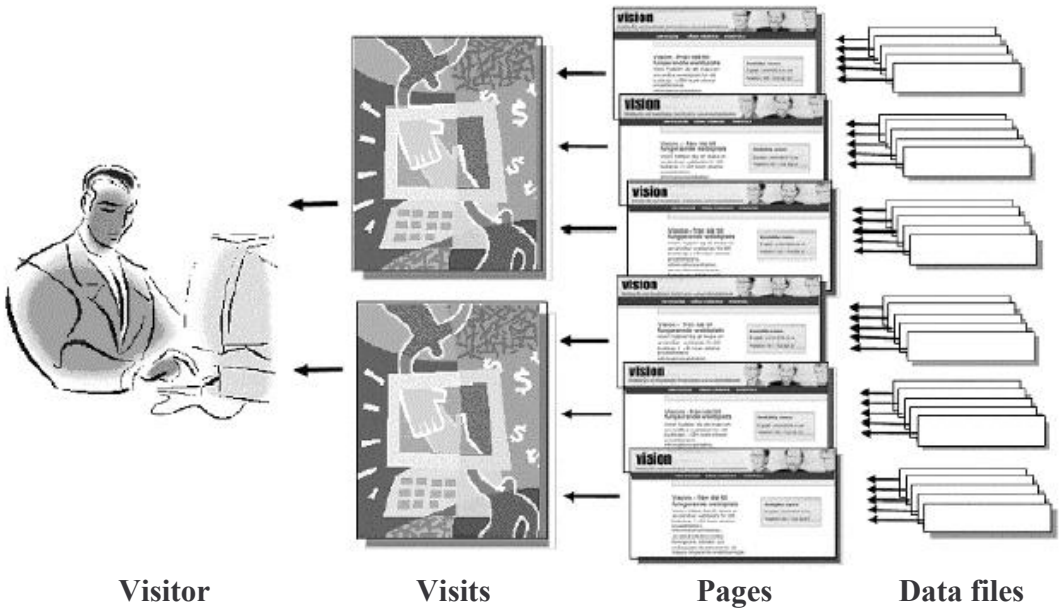
When you talk about visitors you may mean persons, but all a web server has access to is IP-addresses. Between the person that is surfing the web is a web browser and computer, and sometimes a proxy server and firewall. Proxy servers or firewalls can make a number of visitors to get the same IP-address (see picture 1). If you use cookies, you know how many web browsers that have visited your site, but the relation between web browser and visitor is not clear-cut. One person can use a number of browsers, and several persons can use the same browser. This means that persons logging in to a web site would be a more secure way to identify persons, but it is not realistic to demand login on public sites. So how should you define a visitor? One way is to count IP-addresses which is pretty simple but do not say much about the actual visitors. You can count the web browsers that are visiting your site with the use of cookies. Web browser visits are more correct but also more complicated to measure. In addition, there are a lot of search engines that are visiting sites that should not be counted as physical persons.



Picture 1: Overview of how firewalls and proxy servers complicate how to count visitors (Kronman, 2003).

So, what is a web site? Every page is represented by a deliverance of data files that are being logged. Only one of the files contains the pages content, the rest is pictures, java scripts, frames etc.

What is a visit? A visit is a series of files that are being transferred to the same IP-address. After a break of 30 minutes, ends the visit. Here you can also count IP-addresses, with the problem mentioned earlier about proxies. Web browser visits is a series of data files delivered to the same web browser, which demands the use of cookies. (Kronman, 2003) In the picture below (picture 2), one visitor makes two visits to a web site, and gets three pages at each visit. The pages consist of five items each, which means that every page generates five calls to the web server.



Picture 2: Relations between a visitor, a visit, web pages and calls to the server (Kronman, 2003).

When there is awareness about these issues, there is a possibility to make more realistic questions, for example no more questions about persons and places. Instead questions about delivered page views and what parts of the web site that is most used can be answered.

2.3 The need for standards

A number of different systems exist today, using different techniques for traffic measurement. This results in that the data from one system cannot be compared with data from another system. One example of a problem is Yahoo! that claims that their own tracking shows 100 million unique users each month, but the Internet audience measurement company Media Metrix puts the number closer to 40 million. That is sixty million pairs of eyes that can be translated into millions of dollars worth of potential advertising revenues (Regan, 2000). Mattias Inghe (2003) writes a column on Idg.se, in the section "Teknik och Tester". He describes the web statistic situation where everyone can claim that they are the biggest and the best as anarchy. With all the different systems existing today, everyone can find a way to get the figures they want.

As mentioned in the Introduction (see chapter 1), the organisation KIA (Kommittén för Internetannonsering) is working on recommendations for a standard in Sweden. KIA is a so-called "Joint Industry Committee" and consists of advertising companies, web site owners and selling networks that are cooperating on certain issues. In September 2001 KIA presented a document with recommendations for how to measure audience on the Internet. KIA has recently looked over their recommendations, and in October 2003, their new revised recommendations were released (http://www.kiaindex.se/). However, this

does not seem to have had the desirable effect (Inghe, 2003), at least not to this point in time. This shows that there is still work to do.

The problem of web measurement does not exist only in Sweden. In the United States, advertisers want standardized methods to count audiences on Web sites. Different trade groups have tried to set guidelines for online web measurement, and in 2001 the Internet Measurement Initiative has joined forces with the 65-year-old Advertising Research Foundation. This is the latest attempt by advertising trade groups to solve measurement problems that has existed since advertisement on the Web started to grow in the mid 1990:s. (Maddox, 2001)

Novak and Hoffman (1997) bring up another aspect to the problem with web measurements. They are saying that advertising is expected to be a significant source of revenues on the Internet, and is attracting increasing management attention. But because the industry lacks standards for what to measure and how to measure it, the Web is having difficulties being accepted as an advertising medium. The lack of standardization exists on several fronts when it comes to commercial web sites. There are no established principles for measuring traffic on commercial web sites, and there are no standards for measuring consumer response to advertisements. Neither are there a standard for pricing.

The web is a great place for advertisers to display their products and of course they want as many people to see their advertisements. But how do they know how many that is watching? Regan's figures show that there are great differences that cannot be ignored. This clearly shows the problems in this field, since advertisers often pay according to the size of the Web page audience. Of course this leads to insecurity among the advertisers and may lead to the thought that Internet is not the best place for advertising, just as Novak and Hoffman are discussing. Just by changing web measurement system the number of visitors can be doubled (Inghe, 2003), and the company does not need to anything.

3 Literature study

3.1 Techniques to measure web traffic

During our work we have found four completely different techniques for gathering data for web site analysis. They all work different technically, which means that they collect different kinds of information or the same information, but in different ways. Here follows the four major ways of measuring that we have encountered and some of their advantages and disadvantages.

3.1.1 Log file analysis

Server log files have always been seen as the primary source of information for web site traffic and user behaviour. According to Page (1997), the majority of web traffic analysis tools use these log files to see web traffic details.

Every time a file is retrieved from a server, a post in a log file is created. This is done every time a request for a URL comes to the server. To analyse these log files you need an external program or look at them in another program like Microsoft Excel. (Hjelm, 1995, p 111) The software installed on the company's server captures data from the log files for subsequent analysis. The stored data and analytic tools reside within the enterprise. This solution is often more expensive than a web analysis service provider. Although some companies may choose this in-house solution because information about their customers' behaviour may be too strategic too farm out to a third part. (Sweeney, 2001)

There is a possibility for inaccuracy with this method, and that is because log-file data can be infiltrated by "spiders" or "bots", automated programs that accesses web pages, or miss hits because pages are cached (Silber, 2000).

Log file analysis was the first way to gather information about visitors, and most software today uses these log files. However these software are often more expensive then web analysis service providers and there is a risk that you are not getting the correct results. Log files are not able to count the exact number of unique visitors, since every visit is seen as a new visit. The log file analysis software uses an algorithm to calculate the number of unique visitors, and this might not be correct. After reading literature about log files, we have found that there is an inaccuracy of the log file analysis tools.

3.1.2 Web-based statistic services

There are a number of different ASP solutions on the market. ASP stands for Application Service Provider, and this kind of solutions will be referred to as web-based statistic services. When signing up for a web-based statistic service the user will receive a piece of code, which must be inserted on the web page (Geijer, 1999). The code executes on page load and the data is usually stored in a secured database on a separate server, which allows you to use data mining functions in real-time without hurting your Web server's performance.

One advantage with these solutions is that the users do not have to invest in hardware infrastructure or training and the program can be running immediately (Coopee, 2000). They also have another benefit; they can log more information about the surfer's computer, for example the monitor resolution settings and number of colours that are displayable. This is due to the fact that you have pasted a JavaScript into your code that can access information that is never being stored in an ordinary log file (Geijer, 1999).

One disadvantage with the ASP software is the fact that a third party will have access to your maybe closely guarded information, which can mean a security risk (Coopee, 2000).

With an ASP solution you do not need to have the access to your log file to do analysis of your web page, which you may not have if a web hotel hosts your site.

3.1.3 Network-based Web analysis tools

This method is located between the network interface and the lowest level of the host's network code – the best location possible for intercepting HTTP-traffic. This method, also called on-the-wire, gets an unfiltered stream of data which contains all communication between the clients and the server.

This method is collecting data directly from the network and can not only tell if a page was requested, but also if it was successfully delivered.

A traditional protocol analyser is not typically used for measuring performance; they are mostly used for protocol debugging and network testing. (Page, 1997)

3.1.4 Panel/audience measuring

When it comes to panel-based research, the method is to monitor a sample audience and then apply the results for the whole target group (Silber, 2002). A problem here is how to know that the panel is representative of a larger Internet audience. Another fact is that when people are aware of the fact that they are being monitored, they change their behaviour. (Shaw, 2001)

Representative sampling accurately measures only the largest, broadest sites, for which a sufficient sample can be obtained. (Silber, 2000)

This is not a good alternative for measuring web pages traffic for single businesses; it is better for measuring peoples surfing habits and see larger trends.

3.1.5 Combination of methods

We have found that some companies combine different types of methods, for example log file analysis or web statistic systems with audience measuring. One example is Nationalgeographic.com, which uses log data analysis combined with representative measurements from Nielsen/Netratings. (Silber, 2001)

This is not a method in itself, but many companies seem to combine different methods to increase the reliability of their statistics. The figures from one software program can be verified by using software with another measurement technique.

The Swedish web site http://www.idg.se uses three different systems for visitor statistics. One system analyses the log files on the web servers, one system measures every page view in real time and one system is built-in their software, used for handling advertisements on the sites (Inghe, 2003). Inghe says that they are lucky to have all systems showing just about the same figures, but says that not all companies are that lucky.

3.2 Web metrics

"Welcome to Web metrics, where a few numbers can mean many things, and a lot of numbers can add up to very little indeed." This says Debra Judge Silber in an article describing the difficulties with measuring a Web site's success (Silber, 2001). She means that the measurements for your success depend on your strategy for your web site.

"Web metrics" is a phrase that is used today to describe the measurements that is of interest in this area. It is used in many articles and seems to be an accepted expression.

We believe that it is important to analyse the metrics that are used, both to avoid comparing "apples and pears" and to avoid inconsistent data. This can help a business to understand the pros and cons for the measures that are being used, and they can avoid spending time and money on analysing metrics they do not need.

To have a large number of page views or hits is of course a metric in itself, but if a big part of these are automatic reloads or pictures the figure could be irrelevant. A better thing to measure is unique visitors and the amount of new respectively frequent visitors. These are adequate measures for how many new visitors that are added and how many of those that is returning to the site. (Yancy, 2001)

Regan (2002) writes that focus has to shift from quantity to quality. What are the customers getting for they money and how likely are they to come back? Questions like that are not answered by page hits alone. He ends with saying that "when qualitative information about e-commerce sites becomes more widely available and the ratings are finally seen for what they are – a piece of the puzzle, rather than the whole pie -- the statistics will become less important and more valuable." (Regan, 2002).

So, what to measure seems to a big issue in itself and it may not be easy to decide what it is you need to be able to analyse the success of your web site. What to measure seems to be a big issue that has no simple solution, and before you know what to measure it is hard to decide the best way to get it.

3.2.1 Different needs for information

The demand for data from web site analysis depends on who is asking for the information, a technician is asking for different things than the marketing section. For a Webmaster it is important to measure the number of page views, search paths and which browsers the visitors are using. A technician wants information about the band width, error codes that indicates that the server is not functioning at its optimal powers, how long time does it take to deliver web sites and how the balancing of the traffic is between the web servers. A company's marketing section want to know how many of the visitors that was buying and which web sites and ad campaigns were they referred from. (Yancy, 2001)

Mitchell Praver, president of Nationalgeographic.com, confirms the fact that companies may need different kinds of information; "We have many constituencies. For advertisers, there is one set of metrics, for e-commerce there's another set, and for editorial there's another set" (Silber, 2001). This confirms that what you need to measure depends on your needs of information.

The web has moved from being a technology pipe to being a sales channel, says Susannah Patton, and continues that companies need to update their Web measurement strategy with new metrics and analysis tools. However, the new metrics are not clear-out. "There is no standard metric that a company can rely on for its web site. Metrics will be different from company to company", says Randy Souza, analyst at Forrester Research in Cambridge, Mass. (Patton, 2002). It is obvious that different needs for information need different analyses, and it is important that the data that is being analyzed is correct. Otherwise it can lead to wrong or bad decisions.

3.2.2 What information can you get?

There is a large quantity of information that you are able to get from the web-based statistic services about the visitors and their behaviour. Here are some examples:

- The number of hits or visits, every separate item on a page generates a hit when the visitor clicks. For example the page itself generates one hit and each picture on the page will generate on hit
- Page views, the number of HTML pages that are viewed, no matter of how many hits there are on a page
- The most and least popular pages of the site
- Average number of HTML pages per visit
- Average time per visit
- Most common referring sites (which pages has a link to your site)
- Which browsers are used when visiting your site
- Most popular operating systems of your visitors
- Most popular visitor's organizations, for example .com, .org
- Most common countries visitors come from

(Morris, 1998)

There also exists more advanced information, for example for e-commerce. One example is click rate, which is the number of clicks it takes for a potential customer to make a purchase. The fewer clicks the better.

3.2.3 Different types of web sites need different metrics

Susannah Patton, a senior writer for the US magazine CIO that provides information for IT and Business executives, says that the metric that will be most valuable to a company depends on the purpose of the web site. A retail site might focus on conversion rate while a business-to-business might value site reliability and speed. She has divided business web sites into three categories due to their purpose; Business-to-Consumer, Business-to-Business and Content sites.

Business-to-Consumer/ Retail sites

Web site metrics and analytics are especially important to companies that are selling goods on their web site. Here are a few metrics interesting for this kind of site:

- *Net dollar per visitor* - Companies are able to track online customer behaviour to calculate the amount of money earned per visitor on the site
- *Click stream* - Click stream is a broad method for companies to analyse their customers behaviour, for example where do they enter, where do they go and where do they leave?
- *Customer drop-off rates* - Companies can use click stream data to determine why customers leave their site. (Patton, 2002)

Content sites

Content sites, for example media sites, information portals or governmental sites, can use more basic metrics as page views and unique visitors in order to satisfy advertisers. But they need to know that their visitors are satisfied.

- *Loyalty Index* - A company can track a loyalty index by determining how many times a visitor comes to the site each week, month and year.
- *Customer satisfaction* - Click stream data can tell a lot about the visitors' navigational behaviour, but not how the visitor feels while he is clicking around. To see if the visitor has had a positive experience on the web site, companies can use online surveys. (Patton, 2002)

Business-to-business sites

One reason for businesses to use web sites is to cut costs from their supply chain, which means that a businessperson is probably not visiting the site just to browse around. This means that business sites need to think most about their site's performance and how easily and efficiently the customers can buy what they need.

- *Site performance* - The raw performance numbers are perhaps the most important here. One example is how long time it takes to place an order.

- *User efficiency* - Look for pattern and trends in the user behaviour can tell you if visitors are searching for something too often it may bee something wrong with your design.
- *Average time spent on system* - B-to-B Web sites want to help users get on and off their site as quickly and painlessly as possible. "Our goal is to reduce time spent on our site" says Paul Magin, vice president of product development at HIS Engineering. (Patton, 2002)

3.2.4 Different types of analysis needs different types of metrics

Another approach to categorize the metrics is done by Crane, who asks "Why do some e-businesses fail while others prosper?" (Crane, 2003). The answer he gives is that some companies were unable to create a new set of metrics that fundamentally differs from traditional business, but still are intrinsic to e-business. Too few companies understand that Internet is not simply a channel for business, it actually redefines the business itself, and this non-traditional business frontier demands non-traditional metrics to manage it. Traditional profit-and-loss metrics does not paint the whole customer behaviour picture, because they are only measured once; at the time of the sale. In contrast, metrics for e-commerce, so called e-metrics, offers click stream data that can give a rich picture of all behaviour events that lead, or not lead, to a purchase.

Crane presents a framework for how to measure and analyse relevant e-metrics. He divides the possible analyses into different categories. Here follows a summary of his categories:

Quality e-metrics.
These are basic traffic metrics and should be available at least on a weekly basis. Examples on these are:

- Traffic by page and site area: raw traffic data in terms of clicks, visits and users.
- Page leakage percentage: By identifying pages that have a high percentage of visit termination you can better address site design.
- Next click: the pages that are most frequently followed in a direct sequence from a page
- Previous click: shows the pages that occur most frequently directly prior to a page.

Project-centric metrics.
These rely on click stream data and are designed to facilitate online application processes that will result in measurable technological changes. For example the project manager wants to improve a part of the web site, and together with the analytics consultant works to create a set of metrics around the specific page that needs to be improved. These metrics are often disposable after a project is implemented, if they still are seen as important they should be considered for regular publishing as quality e-metrics.

Deep dive metrics.

These metrics are totally customized and designed to answer any of a large range of questions, including root-cause analysis, hypothesis testing, shopping cart analysis, online customer cluster analysis and other data mining applications. Deep dive metrics often need complex statistical analysis on huge amounts of data and often require integration of multiple databases.

Crane ends with mentioning the importance of having a robust online analytics framework to provide quality analytic solutions to problems that would otherwise be invisible and unquantifiable. He says “The ability to make visible and measurable that which is of utmost importance is crucial to e-commerce success and a critical component of e-business survival (Crane, 2003).

Crane and Patton have two different approaches to how to solve the problem with e-metrics. According to Patton, the type of web site decides the need for information. Crane is looking deeper into e-business sites and what information a company needs to be able to do relevant analyses, necessary to run an e-business.

Crane’s approach is easy to understand and it is perhaps quite natural; we do not think companies perform complicated analyses every week. However it is important to be aware of the needs that exist, and we think that all his categories are necessary. Cranes theory can be applied to Patton’s metrics as well; these two approaches do not exclude each other.

3.2.5 How do you know if your metrics are measuring up?

It is important that companies or others that analyse their web site are aware of what they need for their analyses. One idea is that every company or web site owner should ask themselves what they measure, how they do it and why they are measuring the things they do.

The first thing is perhaps to see if the measures are reliable. With reliable we mean that you get the same results over and over again if what you measure does not change. So what evidence is there that the measure that is being used is reliable and measured in a repeatable and consistent way? If the measure/metric is considered to be reliable, the next thing to be considered is whether the measure is valid for the intended purpose. Is it really measuring what it is intended to measure?

Another thing to consider is how sensitive the measure is to changes in the underlying structure. How much does the underlying structure need to change before it is noticeable, and how fast does the metrics change according to changes in the underlying structure?

The cost/benefit aspect is also important. Do the benefits of the measure outweigh all the costs involved in the collection, input, analysis, and reporting of the measure?

How easy is the measure to understand? Can it be easily described and understood by the people who look at it? Perhaps it needs to be reconfigured to be easier and simpler?

Finally - how balanced is the portfolio of measures used to assess the health of the business? A balanced set of measures should include measures about each relevant area of the business. Many measures need to be triangulated, or combined and analyzed with other measures in order to present a useful and complete view.
(http://www.marketingprofs.com/3/perla7.asp)

4 Method

4.1 Quantitative methods

Quantitative research collects data that can be translated into numbers and figures. The result can be worked with statistically, and the result can be possible to generalize. According to Holme and Solvang (1991, p 77-78) quantitative methods should be used when you want to say something about the group a selection is made from and see to what extent something occurs.

4.2 Qualitative methods

Qualitative methods give a greater understanding about a problem area. It gives plenty of information about a few units and increases the understanding of the problem
(Holme & Solvang, 1991, p 78). The purpose with these methods is to understand and analyse the problem as a whole (Patel & Davidson, 1994, p 99).

4.3 Choice of method

Methods are tools to gather information and to know which method to use you need to know the purpose with you want to achieve (Holme & Solvang, 1991, p 76). To be able to verify our hypothesis we needed to collect a large amount of data to compare and analyse. To gather that information we considered quantitative methods to be the best choice. Another reason for choosing quantitative methods is because we wanted to collect data that could be worked with statistically (Patel & Davidson, 1994, p 12).

In this thesis, a group of web-based statistic services was selected for a test. The selected services are not particular interesting in themselves. It is the conclusions that are made from the result that we are interested in.

The purpose of the test was to collect data from these services and compare them with a control program to be able to see if web-based statistic services are reliable, which means that they show the same results as our self developed control program.

We considered that by using quantitative methods we could collect enough information to answer our hypothesis. In addition to the test, we chose to do an interview with a person working with web statistics in a web company. The purpose was to get a better understanding for how to collect and use statistical data about web site visitors in a commercial company.

4.4 The test

4.4.1 Purpose

To try our hypothesis, we chose to do practical tests on several different web-based statistic services. The services registered the traffic from the same web site during a test period. During the same period the control program registered the same things and stored

the result in a database. This result was later compared with the result from the five different web-based statistic services. An analysis of the data made it possible to see whether the statistic services showed an accurate result in counting unique visitors and total number of hits. Since all of these were registering the traffic on the same web page, they should therefore show the same results if they measure correctly.

4.4.2 Collection of data

Data will be collected every week and a final measuring will be done for the whole period. During the same time the control program collects data. The test-period was started on the11: th of March 2003 and went on for almost three weeks forward. The result from the different services will be analysed and compared to each other and to the figures from the control program.

The web page that will be used for the test is http://www.cocktails.nu. It has about 4000 unique visitors each month according to the log files, which will probably be enough to get good measurements. The web site is in English and has visitors from all over the world. A web hotel in Sweden hosts the site.

4.4.3 The facts to measure

There are a lot of things to measure about web site visitors, and what information you need depends on why you need it. To be able to answer our research questions we have chosen some figures that are easy to measure and compare. The things we are going to measure are:

- the total number of pages viewed
- the number of unique visitors

These are basic metrics, and they are also the only things that cannot be manipulated by the user. According to Robin Ericsson, System Administrator at NoName4Us AB, the only things that can be measured for sure are IP-address, page views, hits and unique visitors. This is registered by the server and cannot be modified by the user. The data sent by the web browser, such as type of web browser, referrer, operating system etc. can be set by the user in certain web browsers.

4.4.4 The control program

A series of consecutive and related requests made during a single visit is called a session, and one session is distinguished from the next by a “time-out” period, for example on 30 minutes. If a user does not interact with the web page within the time-out period, the user’s next interaction will start a new session. (Menascé, 2002, page 129)

By using sessions in PHP the control program is able to count the exact number of unique visitors and page views. A visitor accessing a web site is assigned a unique id, the so-called session id. This is either stored in a cookie on the user side or is propagated in the URL. This enables us to count exactly even if the visitor’s web browser does not accept cookies. There is a session time-out for 60 minutes set by the web server.

(http://www.php.net/sessions)

The PHP script is executed every time the page is loaded. By calling the function "`session_start();`" the session starts. The function creates a new session if a session does not already exist. After creating or continuing a session, a check is made whether this is the first time a visitor enters or not. If it is a new user, an entry is made in the table `logg_sessions` in the database. To be able to make the check if it is a new user the script sets a session variable when a new visitor enters.

Each time a page is viewed an entry is made in the table `logg_hits`. By doing this it is possible to count the number of page views. A MySQL database was used to store the results (see appendix III).

To get unique visitors we count the number of posts in the table `logg_sessions`. To get the number of page views we count the rows in the table `logg_hits`.

Also referrers are logged in this script, because we also wanted to include that in the test. Unfortunately we soon discovered that this result could not be compared with the other statistic services, since they display referrers very differently.

For the PHP script code, see appendix III.

4.5 The interview

To verify our hypothesis and to get a better understanding of the problems with web statistics, we made an interview with Robin Ericsson, system administrator at NoName4Us in Malmö. He has been working with the administration of large web sites, for example Cdon.com and viasat.se, for several years and has great knowledge in this area.

We chose to contact him because we have been in contact with him earlier and knew of his interest in and knowledge of web sites. We contacted him and informed him about our work and asked him if he would like to answer a few questions about web statistics, and he agreed. We decided to do an informal telephone interview with some predefined questions, and it was conducted 10 March 2003.

4.6 Delimitations

Due to the circumstance that we did not have access to a web server, web-based statistic services are used in the test. The services had to be free of charge or offering a trial version that we could use.

Five statistic services were tested. The reason for choosing only five was that the performance of the web site must not be deteriorated because of these tests.

As we mentioned earlier we are not going into the usability issues about these services; we are only concentrating on the results they produce.

4.7 The web-based statistic services to test

4.7.1 Selection of the services

The five web-based statistic services to be tested have been chosen by using the search engine Google (www.google.com). Google is an advanced search engine that according to their web page "… uses sophisticated text-matching techniques to find pages that are both important and relevant to your search. For instance, when Google analyses a page, it looks at what the other pages linking to that page have to say about it". (http://www.google.com/intl/sv/help/basics.html)

By the result we got from Google, the first five hits matching the criteria was chosen.
The criteria were:

- the services should be web based because we do not have access to the web server where the test site is situated
- the services should be free of charge or offer a free trial version
- the services should offer similar services, for example number of page views, unique visitors and referrers
- the services should have the ability to create reports for a certain period of time

4.7.1.1 FreeStats

FreeStats is a free service which is advertiser sponsored. This means that a banner advertisement must be placed on sites using FreeStats basic stats package. For those wanting to use FreeStats without a banner advertisement they offer a bannerless upgrade for a price beginning at $8.33 per month. This will allow users to not only use the tracking service with no banner at all, but also give them access to more advanced and detailed stats for their website. (http://www.freestats.com/)

4.7.1.2 GoStats

GoStats provides live web-site usage tracking/reporting & web hit counters. With GoStats professional version you get a customisable display counter (link-back not required) or an invisible counter. Behind the scenes you also get reporting the GoStats powerful professional site statistics analyser. GoStats Free consists of a hit counter, which is an image displayed on your website that tells you and your visitors how many people have seen your website. GoStats also provides a report section that shows you much more information about your visitors. (http://www.gostats.com/)

It can provide you with free stats because it displays banner advertisements on its pages.

4.7.1.3 CQCounter

With this service you place a small image on your web site and they register information about your visitors. This free service does not display any ads on your web site.
(http://www.cqcounter.com/)

4.7.1.4 HitBox

Founded in 1996, WebSideStory pioneered real-time web analytics and is the company that created HitBox. HitBox is the de facto standard for outsourced web analytics and optimisation. With 30 billion being tracked per month it is the industry leader. Clients include British Airways, Cisco Systems, GE Capital, Intel, Mitsubishi Motors, Motorola, Nokia, Northwest Airlines and Sun Microsystems. (http://www.hitbox.com/,)

HitBox Personal is a free service for web site statistics and it is used by 125 000 sites worldwide. (http://www.hitboxcentral.com)

4.7.1.5 HitsLink

Since 1999 Net Applications has been a leading source of tools and utilities for webmasters and eMarketers for the small to medium enterprise. Headquartered in Aliso Viejo California, Net Applications distributes its services through over 2,000 partners and affiliates. It offers business website tools to measure your traffic, promote, monitor, and advertise your site, and HitsLink is one of them. This service is not free, but the company offers a free 30 days trial which we are using. It has two versions, Professional Edition that offers traffic statistics, and Enterprise Editions that besides traffic statistics also offers some additional services for tracking customers. (http://www.hitslink.com/)

5 Result

5.1 The traffic measurement

The five web-based statistic services and the control program were running on the test site for a period of almost three weeks, between 2003-03-11 and 2003-03-30. The traffic was read three times within this time period, 2003-03-16, 2003-03-23 and the final measurement to get the total result at 2003-03-30.

Here are two tables showing the results. Figure 3 below shows a summary of the results regarding unique visitors for the three times of measurement.

	1	2	3
Control program	1 271	2 935	5 050
CQCounter	579	1 420	2 134
FreeStats	392	1 000	1 449
GoStats	2 290	5 532	8 422
Hitbox	358	3 987	5 685
Hitslink	602	1 514	2 244

Figure 3: Unique visitors

In figure 4 you can see the results for the total number of page views for the three times of measurement.

	1	2	3
Control program	4 921	16 493	22 891
CQCounter	3 239	7 736	11 550
FreeStats	1 127	3 027	4 194
GoStats	4 091	9 646	14 205
Hitbox	1 685	4 027	9 250
Hitslink	4 298	10 219	14 872

Figure 4: Page views

5.2 The interview

The interview showed that Robin Ericsson was aware that there are problems about getting correct web statistics. At the company he works at, they use web statistics to monitor the bandwidth and the web site owners use it for statistical purposes. On larger sites the company combined different methods to verify the statistics. He explained what measurable data that could be trusted, which were IP-address, page views, hits and unique visitors. One reason that other information could not be trusted is that the data is sent by the web browser and can be set by the user in certain web browsers. Examples on such information are type of web browser, referrer, operative system etc.

6 Analysis of the result

6.1 The traffic measurement

The analysis of the result from the five services and the control program showed big differences, both between the tested services themselves and compared to the control program.

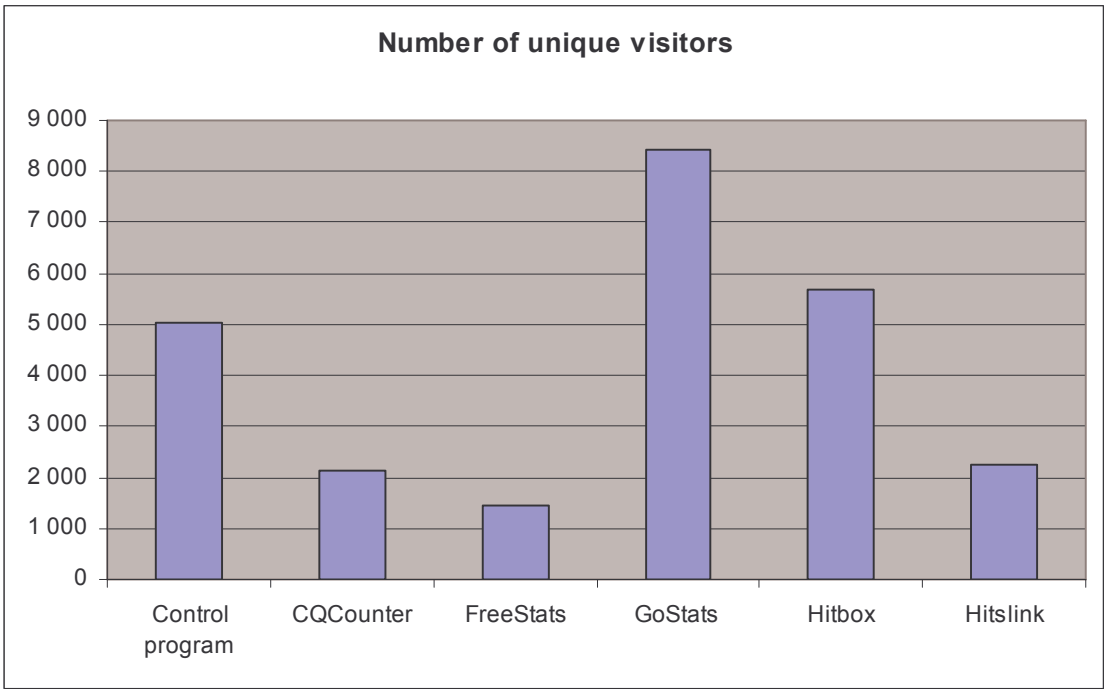


Figure 5: Unique visitors

The number of unique visitors on the web page http://www.cocktails.nu differed a lot between the services in the test. According to FreeStats the number of visitors was only 1 449, but GoStats showed the highest number with totally 8 422 visitors, which is over 500% more (see figure 5). HitBox, with 5 685 visitors, was the service that came closest to the figures from the control program (5 050 visitors), with a difference of +13%.

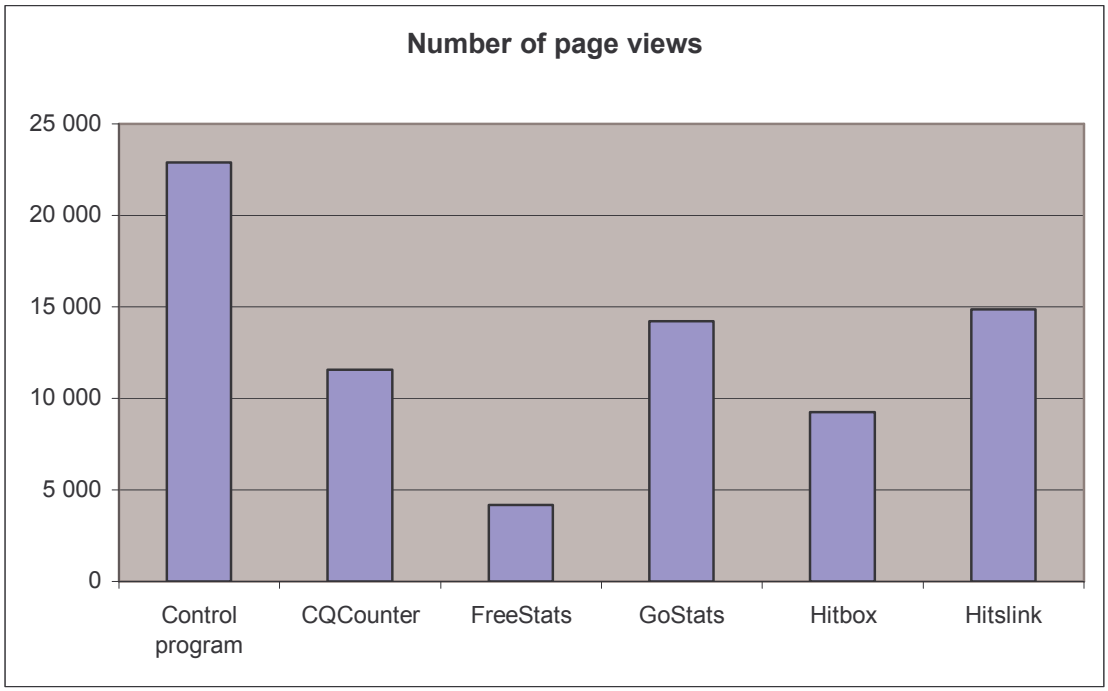


Figure 6: Number of page views

Also when it came to the total number of page views, there was a big difference. Our control program gave us the number of 22 891 page views. All the services in the test showed a lower result. Here too, FreeStats showed the lowest result with only 18% of the actual number of page views. Closest to the control program was Hitslink with 14 872 page views, but even if this was the closest result it only showed 65% of the page views (see figure 6).

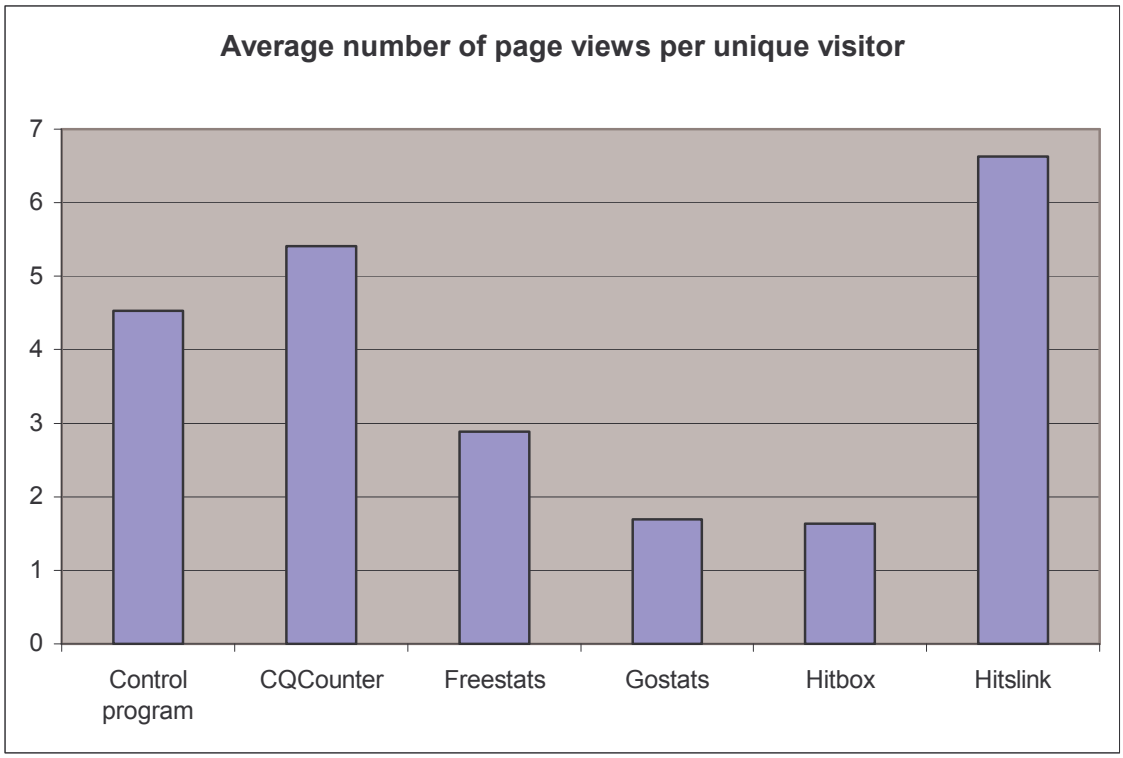


Figure 7: Average number of page views per unique visitor

The analysis of the result from the control program shows that every unique visitor on the test site made about 4.5 clicks each during the test period (see figure 7). This is an interesting figure because it shows how long time the average user spends on the site, which can give a clue about how interesting the user finds it. The longer they stay, the more they like it. But it can also mean the opposite depending on the purpose with the site; perhaps they are clicking around because they cannot find the information they are looking for, which could imply that the site is difficult to navigate. Also here the services showed very different results, from Hitbox's 1.63 to Hitslink's 6.63 page views per visitor.

Using algorithms to calculate unique visitors

The web-based statistic services were not able to count the number of unique visitors, and the reason for this is because they are stateless. This means that every visitor is seen as a new visitor every time he or she enters a page, even if he or she has visited the site recently. Therefore the statistic services have to use some kind of algorithm to be able to estimate the number of unique visitors. (Fletcher, 2002)

From the results in figure 7 you can see how good the different services' algorithms are to calculate unique visitors. As the diagram shows, none of the services could calculate correctly since the relation between hits and unique visitors differs from the control program. This means that even if the number of hits should be measured correctly, they would still not be able to calculate unique visitors correctly.

6.2 Interview

The interview confirmed that there are problems with getting correct information about web site visitors. The interviewee was aware of the problem and the company he worked in had taken actions against it. Their method to get better quality of their statistics was to combine different measurement techniques.

6.3 Conclusion

The research questions we have been trying to answer in this thesis are:

Can web-based statistic services calculate unique visitors correctly?
The relation between total page views and unique visitors is different for every service, which indicates that the algorithms used for calculating unique visitors are not correct. This means that even if they had come to the same result according to total number of page views, the number of unique visitors would still differ.

Is it possible to use web-based services to estimate visiting trends?
If you use the same statistic service over time you can see whether your traffic increases or decreases, but according to the result from our test you cannot trust it to see how big the difference is.

The difference between the services and our control program is not the same over time. For example CQCounter shows only 66% of the page views at the first measurement point. At the second point of measure it showed 47%. This indicates that the services can tell that you have had more visitors from one month to another when you have in fact had less.

To answer the question; yes, you can actually see whether your traffic increases or decreases, but only if there is a significant change. If the service only shows small differences you cannot rely on this data, it may be the service that has calculated incorrect.

Is there one kind of system that is more reliable than others when measuring unique visitors and page views?
The analysis showed that all the services shows great differences compared with our control program. The service that was closest when it came to the number of unique visitors was Hitbox, which showed 13% more visitors. But when it came to the total number of page views, Hitbox only showed 40% of the total page views. The closest according to page views was Hitslink with 65% of the total number of page views.

None of the web-based statistic services showed a result similar to the control program, neither for unique visitors nor total number of page views. Because of this we cannot say that one of the services is better or more reliable than the others.

Our hypothesis: Web-based statistic services do not show an accurate result.
The test showed that there are big differences between the results from the web-based statistic services and that none of them showed an accurate result neither for the total number of page views nor unique visitors according to the control program. This leads us to the conclusion that web-based statistic services do not show an accurate result, which verifies our hypothesis.

7 Discussion

7.1 About the work

As a whole the work with our thesis went well, but of course we encountered some problems. One difficulty was to find literature about this subject. There is not much written about statistic services and the problems in traffic measurement. Another difficulty we encountered was to find services that presented their data in similar ways. Some of the services we found were impossible to compare with others. After finding five services for our test, we started the data collection. It went according to plan except for one thing; we discovered that the data about referring sites could not be compared because they displayed referrers very differently. Therefore we had to exclude referrers from the test. But the fact that we did encounter this problem is a proof in itself of the differences between statistic services, and it shows the problems with comparing data from one software system with data from another.

When you run tests, the result may not be exactly as in real life and one reason can be that you have used a test environment. In our test we used a real live web site, which we think gave results close to reality. The difference between this test and a person using a web-based statistic service is that we used five services whose scripts had to be run after each other.

A combination of quantitative and qualitative methods could have been used to get more understanding about the problem area, for example interviews with users of these services would have been interesting. But due to the limited time it was not possible to also make a larger qualitative investigation. Instead we chose to do one interview that could give us a better understanding of the subject.

We think that the result from the interview confirmed our theory that web-based statistic services cannot be trusted. The interviewee confirmed that he was aware of the situation and that his company had taken steps to verify their statistical information from more than one source.

How representative the result is depends on the selection that has been made. In our case we did not make random samples so it cannot be generalized but we think it would be an unlikely coincidence that we got the only five services that gives an incorrect result.

7.1.1 Possible reasons for differences in the test

There are a few factors that may have had an impact on the result from the services in the test. One is that the different servers that are keeping track of the visitors could have been down for a period of time, which means that visitors have not been registered to that service during that time. The big increase in page views between measure point 1 and 2 on Hitbox indicates that the server may not have been functioning correctly the first week. We have unfortunately not been able to find any information on Hitbox's web page regarding whether the service was up and running during this period.

Another problem that may have occurred is if the visitors have left the site within a very short period of time. For the services to be able to register the visitors and page views, the scripts must be run. At our site, five different scripts have been run after each other and this have taken a few seconds depending on how heavy the traffic was at each service's server at that time. This means that the services that were run first may have registered more page views than the last ones. This would however not have made any big difference, because the scripts are all run within a few seconds.

In our analysis of the result we could not see that the order of the pieces of code on the web site would have made any difference.

7.1.2 Possible reasons for statistic services incorrect results

During our work with this thesis, we have been able to distinguish some different reasons for why the results from traffic analysis software may be incorrect or insufficient. Here are the conclusions we have made:

- The analysing software does not measure correctly. There are a large number of different software systems available, from freeware to expensive systems, and of course they vary in quality.

- There is no standard for measurements; you cannot compare results from different systems. The technique your system is using is affecting the results you get.

- Bad filtering, the analysing tool must be able to sort out visiting robots such as search engines etc. from real persons visiting your site.

- Lack of knowledge, depending on how complicated the analysing software is, a different amount of knowledge is needed to interpret the results in a correct and meaningful way. Solution: go a course, read manuals

- Lack of continuity and routines in the follow up of the analysis. It is important to have a long-term thinking and get to know the analysis tool. In time you will get better to interpret the results from your system, and the trends over time are a good thing in themselves.

- The web site is not built for traffic analysis. One example is frames, which make a web site difficult to analyse because it consists of several pages.

7.2 About the result

7.2.1 Result from the study

The result from the test agrees with the articles that we have found, but the difference between the services was significantly greater then what we had expected. The thing that surprised us mostly was the disability to calculate page views. We expected that page views would be easy to count, but this proved to be wrong.

One of the services, HitsLink, was a trial version. One may think that this service would calculate better because it costs money. It shows a more detailed result and has more functions, but still it does not show an accurate result.

We think that it is important to use the same way for traffic measurement over time, and then you will be able to see the traffic development on the site. If there is a possibility to compare the result from the web-based statistic service to the log file that would probably be a good idea because it could give a hint about how accurate the web statistics are. If users are aware of this problem, it will probably lead to a higher demand for services that follows a possible standard.

We were able to verify our hypothesis by the conclusions we made from the result of the test. Therefore we think that the goal of this thesis; to find out if web-based statistic services are reliable, has been fulfilled.

In our test we found big differences between the results from the different web-based statistic services. That the problem exists is confirmed by the literature we have found. The problem with web statistics is well documented in literature; one example is the measurements for Aftonbladet.se (see chapter 1) and Yahoo.com (see chapter 2.2). However, we have not been able to find any tests of web-based statistics service providers. Our test shows that the problem with web measurements exists there too. This means that changing type of system, for example from traditional log file analysis to web-based statistic services will not give you a more reliable result.

Another confirmation of the problems in this area is the work for standards that is currently going on. Our test proves that you cannot compare data from one software program with data from another, not even if they are using the same method for measuring and the same metrics. One cause for this is perhaps the lack of standards, and we have found articles about the work for standards that is going on, both in Sweden and in the U.S.

7.2.2 The statistics cannot be trusted – so what?

Who is the winner of these inaccurate figures, and who is the loser? It is easy to believe that the users of statistic software want their sites to be popular and have as many visitors as possible. The companies that deliver the statistics want happy customers, so high figures are also wanted here. But the losers in this situation are the adverting companies if they believe in these high visitor statistics. Because if they do, they think that their

advertising campaign reaches more visitors than it actually does, and they probably pay too much for it. So how should they know if they are getting what they pay for? The answer is again, there must be a standard! Already in 1997 Novak and Hoffman brought up this problem, and for what we have seen during our work with this thesis, an acknowledged standard is still missing.

7.3 Suggestions for further studies in this area

Web-based statistic services are likely to expand in the future, and with expansion follow new issues to be handled. Our work is a glance at the problems that exist today, but with the rapid growth of technology, things may change fast in the near future. Suggestions for further research that we have found are:

- An investigation to see if users of these services are aware of the lack of reliability and the fact that the result is not comparable to results from other services. If they are aware of this, how do they handle it?

- It would also be very interesting to analyse the services to try to find out why there are such big variations in the results from these similar services. You could analyse the code that is pasted on the web site and also the algorithms used to calculate unique visitors.

8 References

Books

Fletcher, P, *Practical Web Traffic Analysis*, Glasshaus, 2002

Hjelm, Johan, *Informera på Internet – hur man sätter upp sin egen WWW.* Studentlitteratur, Lund, 1995

Holme, Idar Magne & Solvang, Bernt Krohn., *Forskningsmetodik Om kvalitativa och kvantitativa metoder,* Studentlitteratur, Lund, 1991

Menascé, Daniel A & Almeida, Virgilio A. F., *"Capacity Planning for Web Services"* Prentice Hall PTR, 2002

Patel. R. & Davidson, B., *Forskningsmetodikens grunder.* Studentlitteratur, Lund, 1994

Saarelainen, Kari, *Lokala nät*, Studentlitteratur, Lund, 1996

Articles

Bjurman, Torun., *"Djungel av mätmetoder ger kaos på webben"*, Computer Sweden, 2002-11-11, 2002

Coopee, Todd, *"Going beyond hit counts"*, InfoWorld, 2000-07-17, Volume 22, Issue 29, Pages 45-47, 2000

Crane, Allen S., *"Actionable E-metrics"* Intelligent Enterprise, Vol 6, No 3, 2003-02-01, 2003

Geijer, Erik., *"Webtrends ger avancerad statistik utan program"*, Computer Sweden, 1999-11-18, 1999

Maddox, Kate,. *"Industry groups join to set standards for Web metrics"* B to B, Chicago, 2001-06-11, 2001

Myrén, Karin., "*Webbsverige börjar enas om en standard"*, Computer Sweden, 2003-02-14, 2003

Novak, T.P. and Hoffman, D.L., "*New Metrics for New Media: Toward the Development of Web Measurement Standards*", World Wide Web Journal, winter, 2(1), 213-246, 1997

Page, Bob., "*Network-based analysis tools*", Network World, no:14, 1997

Patton, Susannah., *"Web Metrics That Matter",* Volume 16, issue 4, pages 84-88, CIO Magazine, Framingham, United States, 2002-11-15, 2002

Sweeney, Terry., ”*Site analysis – Web data’s latest buzz”*, InternetWeek, Manhasset, 2001-05-28, 2001

Shaw, Russell., *“Ask every 100th visitor”*, Broadcasting & Cable, volume 131, 2001

Silber, Debra Judge., *“Sizing up Internet benchmarking tools”*, Folio:The Magazine for Magazine Management, issue 1, pages 29-32, 2000

Yancy, Fulton, ”*Webbserveranalys är mer än räkna träffar”*, Nätverk & Kommunikation, nr 17, 2001

Web

Morris, Bruce., *“Software for analyzing your web site traffic”* http://www.webdevelopersjournal.com/columns/analysis.html, 1998, 2003-02-24

Hallström, Lasse., *“Svenska nätpubliken växer fortfarande”*, Internetworld http://www.idg.se/ArticlePages/200302/21/20030221164004_IW/20030221164004_IW.dbp.asp, 2003, 2003-05-01

Kronman, Ulf, *”Statistik från webbplatser – problem och möjligheter”*, 2003-06-02, http://vision.kib.ki.se/portfolio/statistics/pdf/lumano_2003-06-05.pdf, 2003-11-25

Perla, Michael, *“Do Your Metrics Measure Up?”* 2003-06-10, MarketingProfs.com , http://www.marketingprofs.com/3/perla7.asp, 2003-09-30

Regan, Keith., *“Lies, Damned Lies, and Unique Visitors”,* E-commerce Times, 2000-06-21, http://www.ecommercetimes.com/perl/story/3607.html, 2003-05-05

Inghe, Mattias., *“Jakten på den unika besökaren”,* IDG.se, 2003-11-19, http://www.idg.se/ArticlePages/200311/19/20031119093654_IDG.se099/20031119093654_IDG.se099.dbp.asp, 2004-05-07

”Vem sköter företagets webbanalys?”, http://www.publiceringsverktyg.info/1742.html

http://www.google.com/intl/sv/help/basics.html
http://www.kiaindex.se
http://www.passagen.se/funktioner/hjalp/medlemskap/index.shtml
http://www.php.net/sessions

The web page used for the test:
http://www.cocktails.nu

The web pages in the test:
http://www.cqcounter.com/
http://www.freestats.com/
http://www.gostats.com/

http://www.hitbox.com
http://www.hitslink.com/

Previous work
Larsson, Anders & Elofsson, Fredrik, *Analys av webbserverstatistik* Bachelor thesis at Blekinge Institute of Technology, Department of Software Engineering and Computer Science, 2002

Other sources
Ericsson, Robin., System Administrator, NoName4Us AB, telephone interview, 2003-03-10

Appendix I

Glossary

ASP	Abbreviated as ASP, a third-party entity that manages and distributes software-based services and solutions to customers across a wide area network from a central data centre. In essence, ASPs are a way for companies to outsource some or almost all aspects of their information technology needs. They may be commercial ventures that cater to customers, or not-for-profit or government organizations, providing service and support to end users.
Cookie	A message given to a web browser by a web server. The browser stores the message in a text file. The message is sent back to the server each time the browser requests a page from the server.
Data mining	A class of database applications that look for hidden patterns in a group of data that can be used to predict future behaviour. For example, data mining software can help retail companies find customers with common interests. Data mining is increasingly used by marketers trying to distil useful consumer data from Web sites
KIA	Kommittén för InternetAnnonsering. A Swedish organisation that is working on a standard for traffic measurement on the Internet (source: http:www.kiaindex.se)
Log file	A file that lists actions that have occurred. For example, Web servers maintain log files listing every request made to the server.
Page views	A Web page that has been viewed by one visitor. Page views are often used in online advertising, where advertisers use the number of page views a site receives to determine where and how to advertise
PHP	Self-referentially short for PHP: Hypertext Preprocessor, an open source, server-side, HTML embedded scripting language used to create dynamic Web pages. In an HTML document, PHP script (similar syntax to that of Perl or C) is enclosed within special PHP tags.
Traffic	The measurement of the amount of users that visit a Web site.

Unique visitors	When tracking the amount of traffic on a Web site, it refers to a person who visits a Web site more than once within a specified period of time. Software that tracks and counts Web site traffic can distinguish between visitors who only visit the site once and unique visitors who return to the site. Different from a site's hits or page views -- which are measured by the number of files that are requested from a site -- unique visitors are measured according to their unique IP addresses, which are like online fingerprints, and unique visitors are counted only once no matter how many times they visit the site
URL	Abbreviation of Uniform Resource Locator, the global address of documents and other resources on the World Wide Web. The first part of the address indicates what protocol to use, and the second part specifies the IP address or the domain name where the resource is located
Web host	A Web host is in the business of providing server space, Web services and file maintenance for Web sites controlled by individuals or companies that do not have their own Web servers. Many ISPs, such as America Online, will allow subscribers a small amount of server space to host a personal Web page. Other commercial ISPs will charge the user a fee depending on the complexity of the site being hosted.

Source if nothing else is written: http://www.webopedia.com

Appendix II

Interview

Telephone interview with Robin Ericsson, system administrator at NoName4Us AB, Malmö, 10 March 2003.

NoName4Us is a company that builds and hosts web sites. One of the sites is Cdon.se – a site that according to Nielsen/Netratings was the largest e-commerce site in Sweden in April 2003.

(Note: All answers below come from the notes taken during the interview. It is translated from Swedish to English since the interview was held in Swedish.)

Questions asked and answered at the interview:

Does your company use web statistics? If so, for what purpose?
Yes, for bandwidth monitoring and web statistics

What kind of system/ systems does your company use to get statistical information?
Webtrends and Cacti. We use Webtrends primarily to see how much traffic each site has.

How important is the statistics to you?
Medium. It is important that it is functioning correctly, but I do not use it myself. It is handed over to our customers, the web site owners, which work with it.

Are you aware of that the statistics may not be reliable?
Yes, I am aware of the problem with getting correct statistics. On larger sites we use more than one system to verify the statistics.

Which of the measurable data about the visitors can you trust?
The only things that can be measured for sure are IP-address, page views, hits and unique visitors. This is registered by the server and cannot be modified by the user. The data sent by the web browser, such as type of web browser, referrer, operative system etc. can be set by the user in certain web browsers.

Appendix III

PHP script and SQL code

PHP script
Here is the PHP-code used in the control program:

```
[1]   session_start();
[2]   if (!session_is_registered("unique"))
[3]   {
[4]       $query_unique = "INSERT INTO logg_sessions(session) VALUES('"
[5]           .session_id(). "')";
[6]       mysql_query($query_unique);
[7]       session_register("unique");
[8]       $unique = true;
[9]   }
[10]  $query_hits = "INSERT INTO logg_hits(session, page, referrer,
[11]  referrer_dirname, referrer_basename) VALUES('" .session_id(). "',
   '"
[12]  .basename($PHP_SELF). "', '" .$HTTP_REFERER. "', '"
   .dirname($HTTP_REFERER).
[13]  "', '" .basename($HTTP_REFERER). "')";
[14]  mysql_query($query_hits);
```

Database
Here is the SQL-code that was used to create the two tables logg_sessions and logg_hits in the MySQL database:

```
create table logg_sessions(
  id int not null auto_increment,
  primary key(id),
  entered timestamp,
  key(entered),
  session varchar(255)
);

create table logg_hits(
  id int not null auto_increment,
  primary key(id),
  entered timestamp,
  key(entered),
  session varchar(255),
  page varchar(255),
  referrer text,
  referrer_dirname,
  referrer_basename
);
```